RSRE

MEMORANDUM No. 4153

ROYAL SIGNALS & RADAR ESTABLISHMENT

EXPERIMENTS IN MINIMALLY DISTINCT WORD-PAIR DISCRIMINATION USING THE MULTI-LAYER PERCEPTRON

Authors: R K Moore and S M Peeling

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
RSRE MALVERN,
WORCS.

UNLIMITED

RSRE MEMORANDUM No. 4153

R.S.R.E. Memorandum 4153

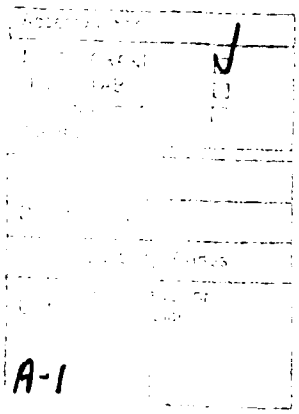
EXPERIMENTS IN MINIMALLY DISTINCT WORD-PAIR DISCRIMINATION USING THE MULTI-LAYER PERCEPTRON

R.K. Moore and S.M. Peeling

April,1988

Abstract

The *multi-layer perceptron* is investigated as a new approach to the automatic discrimination of spoken minimally distinct word-pairs. The choice of the parameters for the multi-layer perceptron is discussed and experimental results are reported. A comparison is made with hidden Markov modelling applied to the same data. The results, for this particular task, show that the discrimination accuracy obtained using the multi-layer perceptron is superior to that attained using hidden Markov modelling.



Contents

List of Figures

List of Tables

1 Introduction

A previous memorandum [3] has discussed the use of the *multi-layer perceptron* (MLP) as a new approach to speech pattern processing and, in particular, to the problem of isolated digit recognition. This memorandum applies the same techniques to the problem of discriminating between minimally distinct word-pairs. Experiments are conducted to show the effect of the choice of the parameters used by the MLP on the experimental results.

A comparison is made with the established technique of hidden Markov modelling applied to the same data. Also, the performance of discrimination simply on the basis of overall energy and length is assessed.

2 The Multi-Layer Perceptron

The units in a multi-layer perceptron are configured in layers such that there is a layer of input units, any number of intermediate layers, and a layer of output units. Connections within a layer or from higher to lower layers are not permitted. Each unit has a real-valued output (between 0 and 1) which is a non-linear function of its total input. For example, the total input, x_j, to unit j is given by:-

$$x_j = \sum_i y_i w_{ij}$$

where w_{ij} is the value of the weighted connection between unit i and unit j.

The output of unit j, y_j, is given by:-

$$y_j = \frac{1}{1 + e^{-x_j}}$$

Thus, given an input pattern, the output pattern can be computed in a single *forward pass* through the network.

If a unit j is an output unit then, for a given target value t_j, the total error E at the output is defined by the following expression:—

$$E = \frac{1}{2} \sum_c \sum_j (t_{jc} - y_{jc})^2$$

where c is an index over input-output pairs.

The learning algorithm minimises E by gradient descent. This involves changing the weights according to the following rule:—

$$\Delta w_{ji}(n+1) = \epsilon \delta_j y_i + \alpha \Delta w_{ji}(n)$$

where Δw_{ji} is the change to be made to the weight on the connection from the ith to the jth unit, ϵ is the learning rate, α is a 'momentum' term and δ is a measure of the local error at unit j.

For an output unit, the error term is given by the expression —

$$\delta_j = (t_j - y_j)\, y_j\, (1 - y_j)$$

and for an internal (hidden) unit the expression is:—

$$\delta_j = \sum_k \delta_k w_{kj}\, y_j\, (1 - y_j)$$

From the foregoing it can be seen that the learning algorithm changes the weights by apportioning the error at the output using a *backward pass* from the output layer to the input layer. This process is termed '*error back-propagation*'.

The effect of the learning algorithm is thus to 'discover' a set of weights which produce an appropriate *non-linear* transformation between input and output. The MLP is thus a powerful technique for deriving high-order internal representations.

A more comprehensive description of the MLP can be found in [3] and [4].

3 Minimal Pair Experiments

3.1 Recorded Speech Data

The data for the experiments described in this memorandum consisted of pairs of minimally distinct spoken words. The minimal pairs used all have the property that their durational characteristics provide an important cue for discrimination. The eleven minimal pairs are shown in Table 1, together with the mnemonics used to refer to them in the later text.

Minimal Pair	Mnemonic
"chip" - "ship"	CSHIP
"cloze" - "close"	CLOZSE
"five" - "fife"	FIVFE
"hard" - "heart"	HARDT
"heard" - "hurt"	HERDT
"killed" - "kilt"	KILDT
"league" - "leak"	LEEGK
"rider" - "writer"	RIDTER
"robe" - "rope"	ROBPE
"seen" - "teen"	STEEN
"wand" - "want"	WONDT

Table 1: Minimal word pairs investigated and their mnemonics.

Twenty examples of each word were recorded. Of these, ten were reserved for training purposes and ten for testing purposes.

3.2 Data Representation

Speech pattern data were obtained by passing speech signals through a 19 channel filter-bank analyser with a 20ms frame rate [1]. The output from each channel corresponded to the amplitude of the signal over a particular frequency band. The data were segmented so that the start and end points of each word were known, and each word was labelled. The words ranged in length between 25 and 44 frames (500ms to 880ms).

3.3 Experiments

Initial experiments were conducted to show the effect of varying the learning rate and momentum term for fixed numbers of hidden units, over a given number of pattern presentations, on the recognition performance of the MLP.

Using the experience gained in these experiments, the learning rate and momentum term were fixed and further experiments were conducted to show the effect of varying the number of hidden units on the recognition performance.

As well as the MLP, the statistical speech recognition technique of hidden Markov modelling (HMM), was applied to the same data.

In view of the specific properties of the minimal pairs used here, discrimination simply on the basis of overall energy and length was also investigated.

4 Software Implementation of the MLP

The MLP program was written in Coral66 and run on a VAX8600. The program was written to allow most of the parameters to be user changeable. The MLP was trained by presenting the complete set of training data (i.e. ten examples of each word) repeatedly. After each set of twenty words the MLP weights were updated. Training continued in this way until the termination criterion was satisfied.

The main output from the program consisted of a data file containing details of all the parameters involved in a particular run, plus the set of weights which had been generated. As the program ran it displayed the error-per-word summed over all the output units. Obviously it wasn't practical to print this error after each pattern presentation so it was only printed after some number of patterns had been presented to the system. Usually, the error was summed over ten sets of twenty pattern presentations, then the average error displayed. In the remainder of this memorandum, the term *presentation* refers to the cycle of presenting twenty words to the MLP and adjusting the weights on the connections using error back-propagation.

5 Experimental Strategy

This section gives greater detail of the strategy used in all the experiments for each of the different techniques applied to the data.

5.1 Discrimination on the Basis of Overall Energy and Length

The main difference between the minimal pairs used in these experiments lies in the durational structure of the words. In view of this, it was important to check that it was not possible to discriminate accurately simply on the basis of overall energy or length differences. The total 'energy' was calculated by summing the filter-bank outputs over the whole word.

The results of this type of discrimination are discussed in Section 7.5.

5.2 Hidden Markov Modelling

In the experiments reported here, a 16-state hidden semi-Markov model HSMM with Gaussian state output probability density functions and non-parametric (Ferguson) state duration probability distribution functions [5] was trained on ten examples of each word

For the results reported here, the same testing files were used by the HSMM and the MLP.

5.3 The Multi-Layer Perceptron

For all the experiments reported here the MLPs had a 19 (channel) x 60 (time frames) input array. Words shorter than 60 frames were padded with silence (zeros) and randomly positioned within the input array. (Hence on repeated presentations the words were not always in the same position in the input array). There were two output units, one for each class. The number of hidden units could be varied. In some cases there were no hidden units, i.e. the input and output units were directly connected. All other experiments involved a single hidden layer of between one and fifty hidden units.

The determination of parameters such as learning rate, momentum scaling term and number of hidden units will now be discussed.

The choice of suitable parameters for any particular experiment is non-trivial since the parameters are dependent on the problem and the MLP configuration. For example, values of the learning rate, ϵ, and the momentum term, α, which are suitable for an MLP with one hidden unit may not be suitable for a system with, say, twenty hidden units. Hence, suitable values for ϵ and α can only be found by experimentation for each configuration. Given the finite timescale for this study, a comprehensive search for the optimum set of parameters was not feasible. Details are given later of the experiments that were conducted in order to find suitable values of ϵ and α for MLPs with one and four hidden units.

Similarly, in order to determine the appropriate number of hidden units it is impractical to conduct an exhaustive search. Experiments were conducted on the effect of using up to fifty hidden units and the results are reported later.

In the initial experiments it was believed that the values of the *start-up weights* might be crucial to the successful convergence of the MLP. (These start-up weights are the small random values which are assigned to the weights on all of the connections before the first

pass through the network). A strategy was therefore evolved in which each experiment was repeated five times with different start-up weights on each occasion. It soon became clear that, in these experiments, this was not necessary provided that the convergence criteria (discussed below) were satisfied.

There are two methods of terminating the training phase for an MLP: the MLP is either presented with a specified number of training examples, or the training continues until the total error E at the output units falls below some pre-defined value. Both strategies were employed but with a further limitation - the MLP so trained must give zero errors when tested on the training data. This criterion was relaxed in some of the experiments when after using five different sets of start-up weights the MLP still did not give zero errors on the training data.

Rumelhart et al [4] state that setting the target outputs to be 0 or 1 encourages the weights on the connections to become infinitely large. This is not a problem which has materialised in these experiments. The target outputs were always set to 0 or 1.

6 Computational Requirements

Full details of the computational requirements of the MLP used for isolated digit recognition can be found in [3]. The same conclusions apply to the MLP used for minimal word pair discrimination; the MLP is more computationally expensive than the HSMM during the training phase; but the situation is reversed during the testing phase.

7 Results

7.1 Learning Rate and Momentum Term for a Fixed Number of Pattern Presentations

Initial experiments were conducted to investigate the effect of different values of the learning rate, ϵ, and momentum scaling term, α. For these experiments MLPs with one or four hidden units were used. Each experiment used five different sets of start-up weights and nine different pairs of ϵ and α values. These pairs of values involved ϵ taking the values 0.1, 0.25 and 0.4 with α values of 0.25, 0.5 and 0.75, in all possible combinations for ϵ and α. Each MLP with one hidden unit was trained on 1000 pattern presentations, i.e. it was shown ten examples of words from each of the two classes 1000 times. For the MLP with four hidden units, training involved 400 pattern presentations. The weights were updated after each set of 20 words. The MLP was then used to recognise the training and test data sets.

Graphs showing typical results, from both training and test sets, for both one and four hidden units can be found in Appendix A, together with tables giving the complete results for the eleven individual minimal pairs.

A summary of the results for one hidden unit is shown in Tables 2 and 3, and for four hidden units in Tables 4 and 5. These are the "best" results for each pair of ϵ and α in

ϵ	α		
	0.25	0.50	0.75
0.10	0	0	0
0.25	0	2	11
0.40	6	7	13

Table 2: Total errors from 11 minimally distinct word pairs with a 220 word training set for MLPs with 1 hidden unit and ϵ and α as shown after 1000 pattern presentations

ϵ	α		
	0.25	0.50	0.75
0.10	10	10	10
0.25	21	16	23
0.40	15	17	24

Table 3: Total errors from 11 minimally distinct word pairs with a 220 word test set for MLPs with 1 hidden unit and ϵ and α as shown after 1000 pattern presentations.

ϵ	α		
	0.25	0.50	0.75
0.10	0	0	0
0.25	0	0	0
0.40	0	0	6

Table 4: Total errors from 11 minimally distinct word pairs with a 220 word training set for MLPs with 4 hidden units and ϵ and α as shown after 400 pattern presentations.

ϵ	α		
	0.25	0.50	0.75
0.10	7	6	7
0.25	5	5	9
0.40	5	4	13

Table 5: Total errors from 11 minimally distinct word pairs with a 220 word test set for MLPs with 4 hidden units and ϵ and α as shown after 400 pattern presentations.

the sense that they come from the MLPs which gave the smallest number of errors on the training set. (When more than one MLP gave the same number of errors the results quoted come from the one with the smallest error-per-pattern after 1000 or 400 presentations).

From Tables 2 and 3 it is clear that the best training and test results for experiments using an MLP with one hidden unit came from using ϵ=0.1 with any of the three values of α. From Tables 4 and 5 it is clear that all the results obtained from an MLP with four hidden units are very similar; apart from those obtained by using ϵ=0.4 with α=0.75.

As with the isolated digit experiments, reported in [3], the results obtained showed that the strategy of using five different sets of start-up weights per experiment was unnecessary. All further experiments used just one set of start-up weights unless the MLP failed to converge, i.e. it did not give zero errors on the training set; in that circumstance another set was used. Again, an overall limit of five different sets of start-up weights per experiment was applied. In these experiments it was very rare for a run not to converge. Only in the case of zero hidden units were several sets of start-up weights required and even then it was only for a couple of word pairs.

7.2 Termination Criteria

As mentioned previously, the training phase can be terminated either after a pre-determined number of pattern presentations, or when the error-per-pattern falls below some pre-specified value. In view of the significant improvement in recognition accuracy obtained by using the second criterion in the isolated digit experiments, it was decided to investigate the effect in these minimal pair experiments.

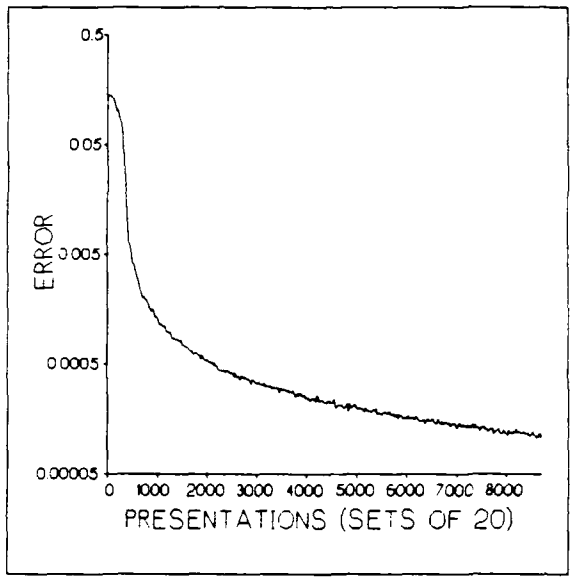


Figure 1: Graph showing the behaviour of the error-per-pattern at the output units for an MLP with one hidden unit, trained on the minimal pair CSHIP.

As an initial test, a subset of the data was used; specifically the minimal pairs CSHIP, HARDT and KILDT. This subset was chosen since experience showed that KILDT produced very few recognition errors, HARDT gave many recognition errors and CSHIP was about average. Experiments were conducted using an MLP with both one and four hidden units and with ϵ=0.1, α=0.5. Training progressed in stages so that the recognition accuracy on both training and test sets could be assessed (the set of weights produced at the end of each stage was used as the starting point for the next stage).

The error at the output units for an MLP with one hidden unit, trained on the minimal pair CSHIP, is shown in Figure 1. (Similar graphs for the other three minimal pairs can be found in Appendix B). For comparison purposes, a graph showing the error at the output units for an MLP used in the isolated digit experiments is shown in Figure 2. The overall behaviour of the error is similar in both graphs. However, in the minimal pair case, many more pattern presentations are needed before the error begins to flatten out.

The test set errors for these three minimal word pairs at each of the various stages are

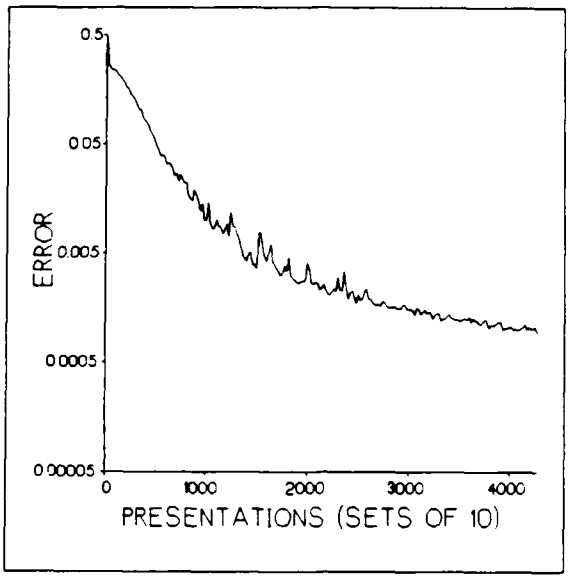


Figure 2: Graph showing typical behaviour of the error-per-pattern at the output units for an MLP in the isolated digit experiments.

shown in Table 6. (In all the cases there were no errors on the training set.)

From the results in Table 6 and the graphs showing the behaviour of the error (Figures 1 and 2), it is clear that the choice of termination criterion for the minimally distinct word pairs is not as obvious as it was for the isolated digits. Looking at the graphs of the error at the output units (in Appendix B) it can be seen that a large number of iterations are needed before there is any flattening out. Whilst the ideal solution is to terminate once the error has flattened out, and preferably with as few test-set errors as possible, it must be borne in mind that extra pattern presentations are computationally expensive. For the one hidden unit case, 1000 iterations take about twelve minutes on a VAX8600, and in the case of four hidden units the figure is about twenty minutes. Bearing all these factors in mind the termination criteria chosen was to iterate until the output error was less than 0.0005 or 0.00025. All remaining results will state which termination criterion has been used.

No. of Hidden	Termination Criterion	Average no. of Presentations	Total Test Set Errors
1	0.01	600	5
1	0.001	1240	8
1	0.0005	2000	7
1	0.00025	3480	8
1	0.0001	7410	7
1	0.00009	8230	10
4	0.01	240	4
4	0.001	820	3
4	0.0005	1250	2
4	0.00025	2650	3
4	0.0001	4530	1
4	0.00009	5270	2

Table 6: Effect of using different termination conditions for MLPs with one and four hidden units – total errors from 60 word test set.

7.3 Learning Rate and Momentum Term for a Variable Number of Pattern Presentations

The sets of experiments using the nine pairs of ϵ and α values were repeated, but terminating when the error fell below 0.0005 or 0.00025. Also, as well as using one and four hidden units, experiments were conducted using MLPs without any hidden units. Full details of the results can be found in Appendix C. A summary of the test set results are shown in Tables 7 and 8 for zero hidden units; in Tables 9 and 10 for one hidden unit; and in Tables 11 and 12 for four hidden units. Except in the cases indicated, there were no errors on the training set.

ϵ	α		
	0.25	0.50	0.75
0.10	9*	5	8
0.25	10	11	13
0.40	10	11	17

*some runs had not converged

Table 7: Total errors from 11 minimally distinct word pairs with a 220 word test set for MLPs with zero hidden units, ϵ and α as shown (learning terminated when error<0.0005).

ϵ	α		
	0.25	0.50	0.75
0.10	6*	10	9
0.25	9	7	13
0.40	10	11	13

*some runs had not converged

Table 8: Total errors from 11 minimally distinct word pairs with a 220 word test set for MLPs with zero hidden units, ϵ and α as shown (learning terminated when error<0.00025).

For the zero hidden unit case the results from the different termination criteria are very similar. There is only a 6% improvement in recognition performance from using the smaller termination condition.

ϵ	α		
	0.25	0.50	0.75
0.10	11	6	11
0.25	12	14	12
0.40	19	12	13*

*one run did not give zero errors on training

Table 9: Total errors from 11 minimally distinct word pairs with a 220 word test set for MLPs with one hidden unit, ϵ and α as shown (learning terminated when error<0.0005).

ϵ	α		
	0.25	0.50	0.75
0.10	7	13	18
0.25	10	12	15
0.40	12	12	13

Table 10: Total errors from 11 minimally distinct word pairs with a 220 word test set for MLPs with one hidden unit, ϵ and α as shown (learning terminated when error<0.00025).

In the one hidden unit case there is no improvement from continuing the learning process - in fact there is a 3% decrease in recognition accuracy.

ϵ	α		
	0.25	0.50	0.75
0.10	4	3	5
0.25	7	4	9
0.40	8	8	13

Table 11: Total errors from 11 minimally distinct word pairs with a 220 word test set for MLPs with four hidden units, ϵ and α as shown (learning terminated when error<0.0005).

ϵ	α		
	0.25	0.50	0.75
0.10	6	5	3
0.25	6	5	8
0.40	6	4	9

Table 12: Total errors from 11 minimally distinct word pairs with a 220 word test set for MLPs with four hidden units, ϵ and α as shown (learning terminated when error<0.00025).

The only case in which there was a marked improvement from using the 0.00025 termination condition was with four hidden units. Here continuing the learning process resulted in 13% fewer recognition errors.

Since over all the experiments the minimum number of recognition errors came from using ϵ=0.1 with α=0.5 these values were used in all the remaining experiments. Furthermore, in all cases these minima resulted from terminating the learning process when the error fell below 0.0005. Hence all remaining experiments use this condition for terminating the learning process.

7.4 Number of Hidden Units

There is no simple way to decide how may hidden units are necessary to solve a specific problem. However, it is known that two hidden layers are sufficient [2].

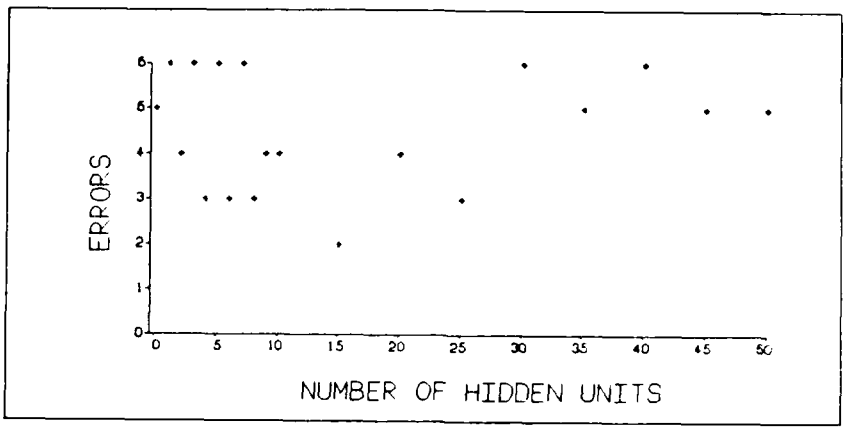


Figure 3: Test set errors versus number of hidden units over eleven minimally distinct word pairs using MLPs with ϵ=0.1, α=0.5.

In order to assess the effect of the number of hidden units on the recognition performance of the MLP ϵ and α were set at 0.1 and 0.5 respectively whilst the number of hidden units was varied. Beginning with zero hidden units the number was increased by one until there were ten hidden units in the system. After this, the hidden units were increased by five each time up to a maximum of fifty units. In all cases the MLPs used gave no errors on the training data. The effect of the increase in hidden units on the test set recognition accuracy is shown in Figure 3.

The number of pattern presentations involved is shown in Figure 4.

From Figures 3 and 4 it can be seen that the best test set recognition is obtained from an MLP with fifteen hidden units. Generally in these experiments, the more hidden units that were used the fewer pattern presentations were needed to train the system. However, the decrease is very slight after the number of hidden units is increased beyond twenty.

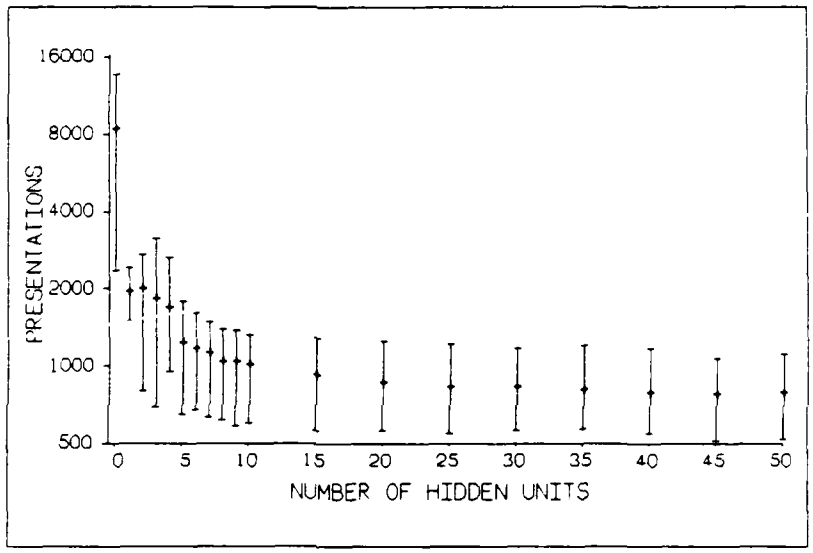


Figure 4: Pattern presentations versus number of hidden units over eleven minimally distinct word pairs using MLPs with ϵ=0.1, α=0.5. The error bars show the range of the presentations in each case. The mean number of presentations is shown by the horizontal line within the bar.

7.5 Comparison with Discrimination on the Basis of Overall Energy and Length

The graphs of overall energy versus length for all the minimal word-pairs described in this memorandum, are shown in Appendix D. The summary of the results obtained is shown in Table 13, together with the 'best' and 'worst' performances obtained from fully trained MLPs. From previous results, the worst fully trained performance is from an MLP with a single hidden unit and ϵ=0.4, α=0.25. The best performance came from an MLP with fifteen hidden units and ϵ=0.1, α=0.5. Note that the errors quoted are from both test and training sets, but all the MLPs gave no errors on the training sets.

Word Pair	Overall Energy/Length	'Worst' MLP	'Best' MLP
CSHIP	8	2	0
CLOZSE	1	0	0
FIVFE	3	1	0
HARDT	1	7	0
HERDT	3	1	0
KILDT	0	0	0
LEEGK	2	0	0
RIDTER	0	2	0
ROBPE	0	4	1
STEEN	5	2	1
WONDT	0	0	0
Total	23	19	2

Table 13: Result of minimal pair discrimination on the basis of overall energy and length as compared to that obtained using MLPs.

From Table 13 it can be seen that only four minimal pairs, KILDT, RIDTER, ROBPE and WONDT, show clear separations between the two word classes using the overall energy/length criteria. It is worth noting that although KILDT and WONDT have consistently given very few recognition errors with all the MLPs used here, neither of the other two pairs stand out from the rest.

7.6 Comparison with Hidden Markov Modelling Results

Table 14 shows a comparison of the total test set errors from each of the eleven word pairs for an MLP with fifteen hidden units (ϵ= 0.1, α=0.5) trained until the error was less then 0.0005, as compared to those from a HSMM.

Word Pair	MLP Errors	HSMM Errors
CSHIP	0	3
CLOZSE	0	0
FIVFE	0	5
HARDT	0	1
HERDT	0	1
KILDT	0	1
LEEGK	0	0
RIDTER	0	0
ROBPE	1	0
STEEN	1	1
WONDT	0	0
Total	2	12

Table 14: Results: errors from 20 word test sets.

From the results in Table 14 it can be seen that the performance obtained from using an MLP with fifteen hidden units is much better than that from using a 16-state HSMM.

8 Conclusions

Experience suggests that for this particular task, unlike the problem of isolated digit recognition, the choice of learning rate and momentum term are not crucial. Neither is the number of hidden units.

The choice of termination criterion is non-trivial for this task. The criterion finally used in this study was a compromise between recognition performance and computer time, bearing in mind the behaviour of the error at the output units

Even the worst fully trained MLP performance is better than that obtained using discrimination on the basis of overall energy and length.

It is clear that for the task of discriminating between the minimally distinct word pairs described in this memorandum, MLPs are capable of a level of performance superior to that obtained by using HSMMs. This is not really surprising since HSMMs are concerned with modelling the words accurately, whilst MLPs are concerned with discriminating between them

References

[1] J.N. Holmes. The JSRU Channel Vocoder. *Proc. IEE*, 127 Pt.F(1):53-60, 1980.

[2] I.D. Longstaff and J.F. Cross. *A Pattern Recognition Approach to Understanding the Multi-layer Perceptron*. Memo 3936, Royal Signals and Radar Establishment, 1986.

[3] S.M. Peeling and R.K. Moore. *Experiments in Isolated Digit Recognition Using the Multi-layer Perceptron*. Memo 4073, Royal Signals and Radar Establishment, 1987.

[4] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning Internal Representations by Error Propagation. In D.E. Rumelhart and J.L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*, MIT Press, 1986.

[5] M.J. Russell and A.E. Cook. Experimental Evaluation of Duration Modelling Techniques for Automatic Speech Recognition. *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing*, :2376-2379, 1987.

Appendix A Effect of Varying Learning Rate and Momentum Over a Fixed Number of Pattern Presentations

The tables here show the errors obtained for the 11 minimally distinct word pairs using an MLP with either one or four hidden units. In the one hidden unit case the MLP saw each of the twenty words 1000 times. For the four hidden unit case there were 400 presentations. The words were positioned randomly within the input array. Each word pair is identified as in the main body of the memorandum.

The results are shown from using ϵ=0.1, 0.25 and 0.4 with α=0.25, 0.5 and 0.75.

There were five runs per word pair for each experiment but only the 'best' results are shown in the tables. Graphs are shown for the results over all the runs for ϵ=0.1, α=0.5 for both one and four hidden units. These are typical and demonstrate the consistency of the results over the five runs.

Word Pair	ϵ=0.1			ϵ=0.25			ϵ=0.4		
	α			α			α		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CSHIP	0	0	0	0	0	0	0	0	5
CLOZSE	0	0	0	0	0	0	0	0	0
FIVFE	0	0	0	0	0	2	1	0	1
HARDT	0	0	0	0	2	6	2	3	1
HERDT	0	0	0	0	0	1	2	3	3
KILDT	0	0	0	0	0	0	0	0	0
LEAGK	0	0	0	0	0	0	0	0	0
RIDTER	0	0	0	0	0	2	1	1	1
ROBPE	0	0	0	0	0	0	0	0	0
STEEN	0	0	0	0	0	0	0	0	2
WONDT	0	0	0	0	0	0	0	0	0

Table 15: Errors from the 20 word training set for the eleven minimally distinct word pairs using one hidden unit.

Word Pair	ϵ=0.1			ϵ=0.25			ϵ=0.4		
	α			α			α		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CSHIP	2	2	1	1	1	2	1	2	6
CLOZSE	0	0	0	0	0	0	0	0	0
FIVFE	0	0	0	4	0	3	2	2	2
HARDT	5	5	5	6	3	6	5	5	3
HERDT	1	1	1	6	4	5	4	3	4
KILDT	0	0	0	0	0	1	0	0	0
LEAGK	0	0	0	0	0	0	0	0	2
RIDTER	1	1	1	2	2	3	1	4	2
ROBPE	0	0	1	1	5	2	1	0	1
STEEN	1	1	1	1	1	1	1	1	3
WONDT	0	0	0	0	0	0	0	0	1

Table 16: Errors from the 20 word test set for the eleven minimally distinct word pairs using one hidden unit.

Word Pair	ϵ=0.1			ϵ=0.25			ϵ=0.4		
	α			α			α		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CSHIP	0	0	0	0	0	0	0	0	3
CLOZSE	0	0	0	0	0	0	0	0	0
FIVFE	0	0	0	0	0	0	0	0	0
HARDT	0	0	0	0	0	0	0	0	2
HERDT	0	0	0	0	0	0	0	0	1
KILDT	0	0	0	0	0	0	0	0	0
LEAGK	0	0	0	0	0	0	0	0	0
RIDTER	0	0	0	0	0	0	0	0	0
ROBPE	0	0	0	0	0	0	0	0	0
STEEN	0	0	0	0	0	0	0	0	0
WONDT	0	0	0	0	0	0	0	0	0

Table 17: Errors from the 20 word training set for the eleven minimally distinct word pairs using four hidden units.

Word Pair	ϵ=0.1			ϵ=0.25			ϵ=0.4		
	α			α			α		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CSHIP	1	1	2	2	2	2	1	1	2
CLOZSE	0	0	0	0	0	0	0	0	0
FIVFE	0	0	0	0	0	1	1	1	1
HARDT	3	3	4	2	2	2	1	1	3
HERDT	1	1	0	0	0	1	0	0	4
KILDT	0	0	0	0	0	0	0	0	0
LEAGK	0	0	0	0	0	0	0	0	0
RIDTER	1	0	0	0	0	0	0	0	0
ROBPE	0	0	0	0	0	0	0	0	0
STEEN	1	1	1	1	1	3	2	1	3
WONDT	0	0	0	0	0	0	0	0	0

Table 18: Errors from the 20 word test set for the eleven minimally distinct word pairs using four hidden units.

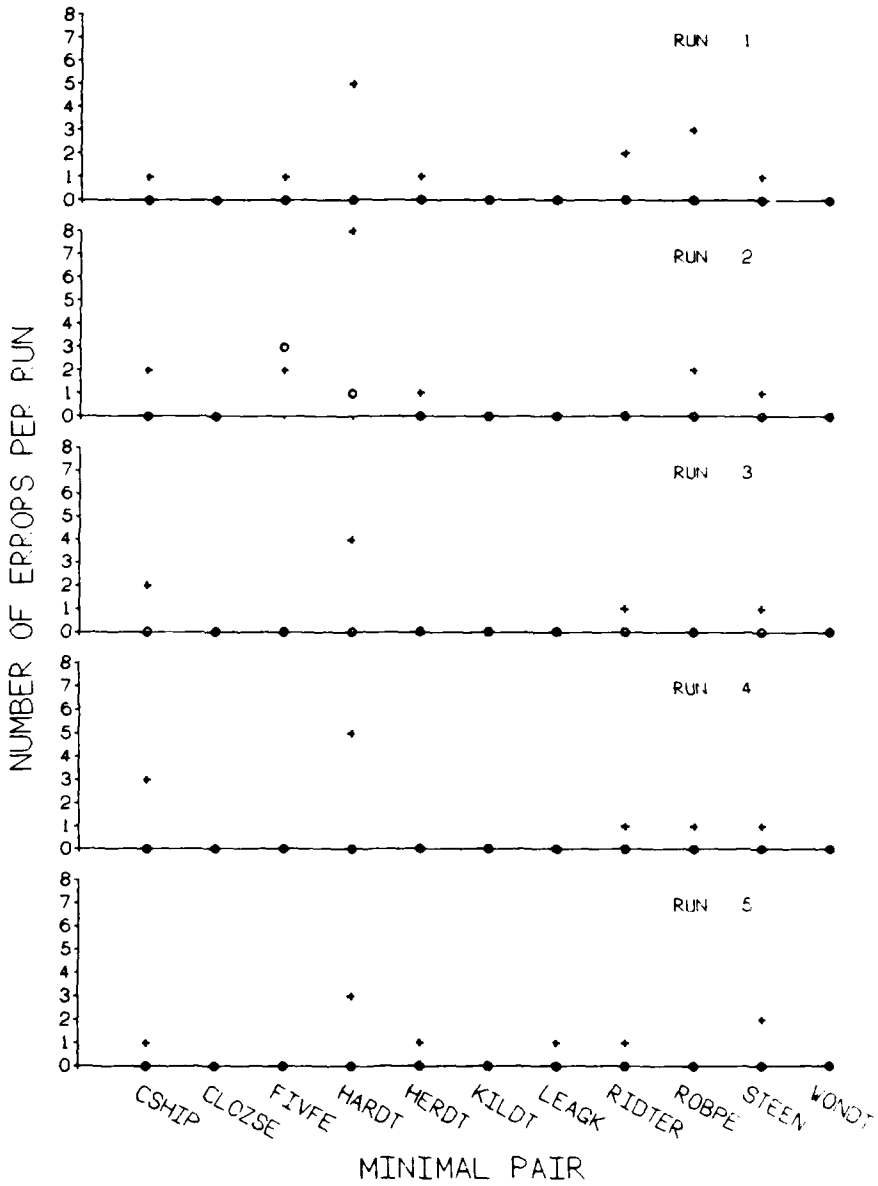


Figure 5: Errors after 1000 pattern presentations from MLP's with one hidden unit, ϵ=0.1 and α=0.5. Circles represent training set errors and crosses represent test set errors, both over 20 words.

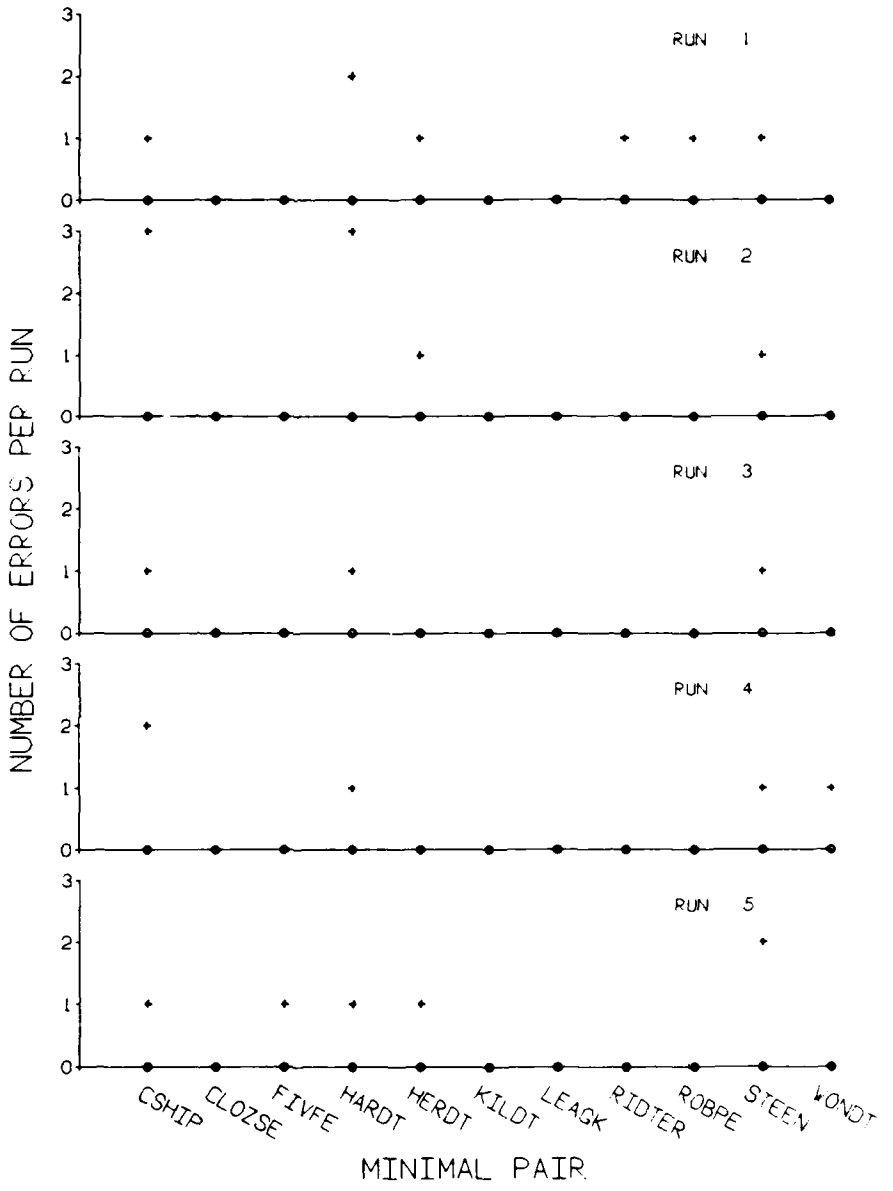
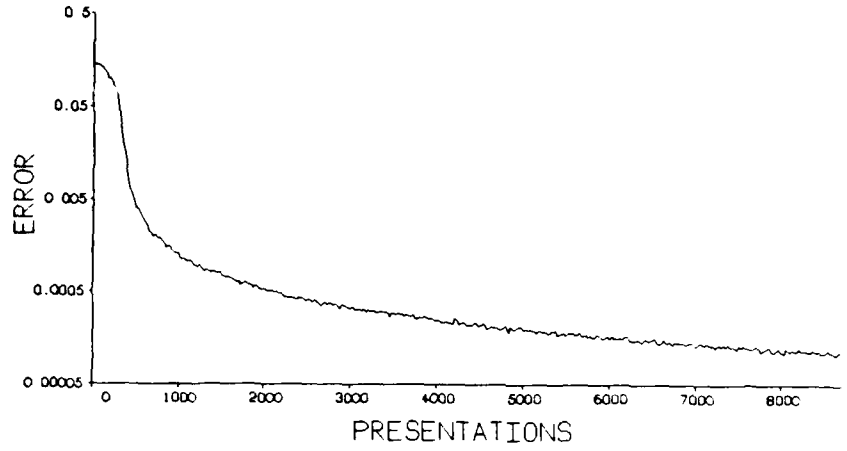


Figure 6: Errors after 400 pattern presentations from MLP's with four hidden units, ϵ=0.1 and α=0.5. Circles represent training set errors and crosses represent test set errors, both over 20 words.

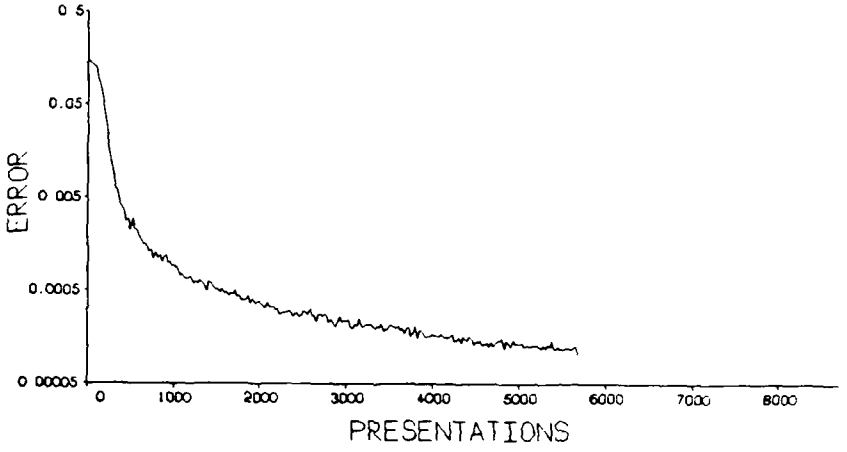
Appendix B The Error at the Output Units as a Function of the Number of Pattern Presentations

The graphs here show the behaviour of the error at the output units as the number of pattern presentations is increased.

Results are shown for MLPs with both one and four hidden units for the word pairs CSHIP, HARDT and KILDT. In all cases ϵ=0.1 and α=0.5.

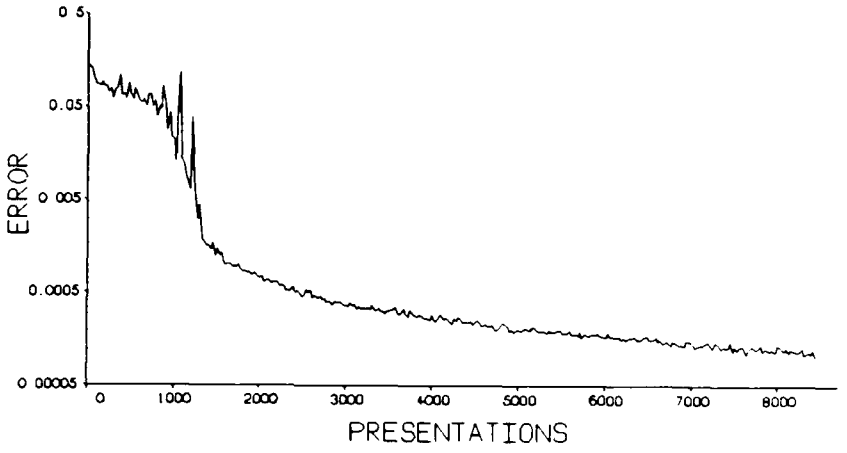


(a) One Hidden Unit

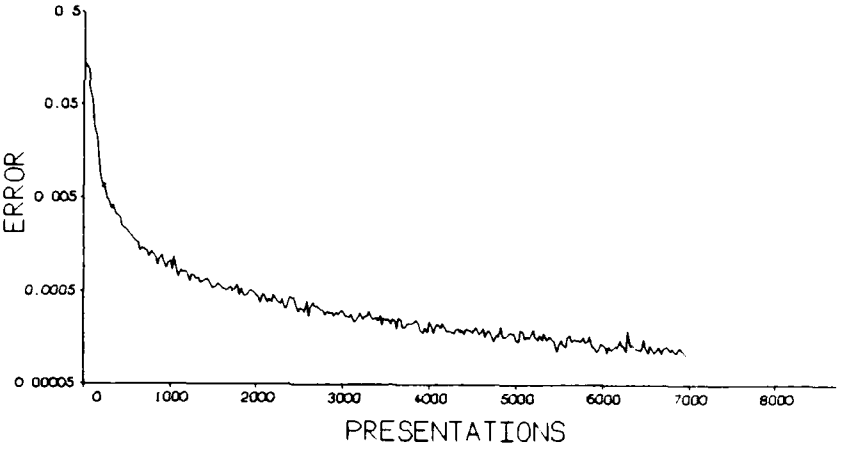


(b) Four Hidden Units

Figure 7: Error at output units, as a function of the number of pattern presentations, for word pair CSHIP with ϵ=0.1, α=0.5 and one hidden unit in (a) and four in (b).

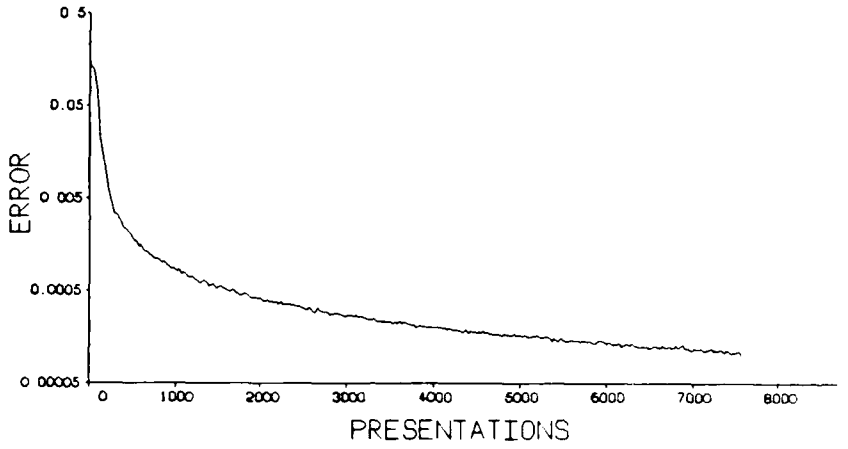


(a) One Hidden Unit

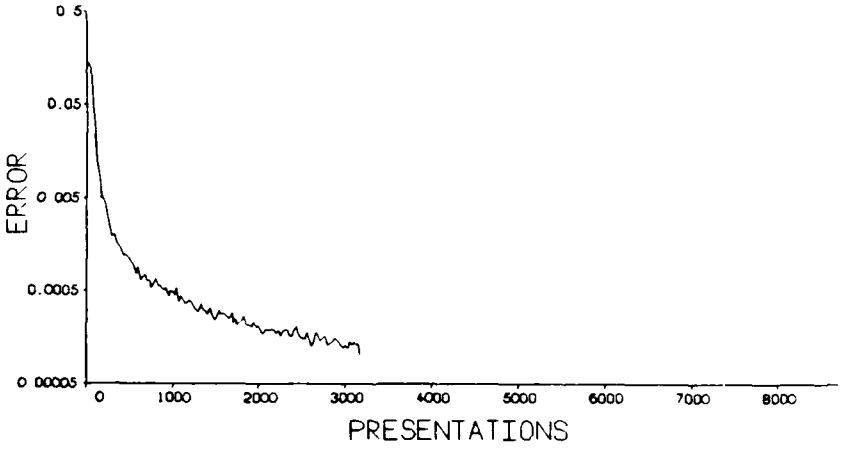


(b) Four Hidden Units

Figure 8: Error at output units, as a function of the number of pattern presentations, for word pair HARDT with ϵ=0.1, α=0.5 and one hidden unit in (a) and four in (b).



(a) One Hidden Unit



(b) Four Hidden Units

Figure 9: Error at output units, as a function of the number of pattern presentations, for word pair KILDT with ϵ=0.1, α=0.5 and one hidden unit in (a) and four in (b).

Appendix C Effect of Varying Learning Rate and Momentum Using Error per Pattern as the Termination Criterion

The tables here show the errors obtained for the 11 minimally distinct word pairs using an MLP with zero, one or four hidden units. In each case, two termination criteria were used. The first result is from terminating the learning process when the error feli below 0.0005; and the second from when the error fell below 0.00025.

The words were positioned randomly within the input array. Each word pair is identified as in the main body of the memorandum.

The results are shown from using ϵ=0.1, 0.25 and 0.4 with α=0.25, 0.5 and 0.75.

If the learning process did not result in zero errors on the training data then up to four more sets of start-up weights were tried. If this still did not result in zero errors then the result quoted comes from the run producing the smallest number of errors on the training data.

Word Pair	ϵ=0.1			ϵ=0.25			ϵ=0.4		
	α			α			α		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CSHIP	1*	1	2	2	1	2	1	1	3
CLOZSE	0	0	0	0	0	0	0	0	0
FIVFE	0*	0	0	1	0	0	1	1	1
HARDT	4*	2	2	4	7	5	5	5	5
HERDT	1*	0	1	0	0	1	1	1	1
KILDT	0	0	0	0	0	0	0	0	0
LEAGK	0	0	0	0	0	0	0	0	0
RIDTER	1	0	0	0	0	2	0	0	2
ROBPE	1	1	1	1	2	2	1	2	2
STEEN	1	1	2	1	1	1	1	1	2
WONDT	0	0	0	1	0	0	0	0	1

*had not converged so stopped after 15000 presentations

Table 19: Test set errors, from 20 words, for the eleven minimally distinct word pairs using zero hidden units and trained until error<0.0005.

Word Pair	ϵ=0.1			ϵ=0.25			ϵ=0.4		
	α			α			α		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CSHIP	0*	1	1	1	1	2	1	0	1
CLOZSE	0	0	0	0	0	0	0	0	1
FIVFE	0*	0	0	1	0	0	1	1	0
HARDT	3*	4	4	2	4	8	4	6	6
HERDT	0*	1	0	0	0	1	1	1	0
KILDT	0	0	0	0	0	0	0	0	1
LEAGK	0	0	0	0	0	0	0	0	0
RIDTER	1	1	1	1	0	1	1	1	2
ROBPE	1	1	2	2	1	0	1	1	1
STEEN	1	1	1	1	1	1	1	1	1
WONDT	0	1	0	1	0	0	0	0	0

*had not converged so stopped after 30000 presentations

Table 20: Test set errors, from 20 words, for the eleven minimally distinct word pairs using zero hidden units and trained until error<0.00025.

Word Pair	ϵ=0.1			ϵ=0.25			ϵ=0.4		
	α			α			α		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CSHIP	2	1	2	2	3	2	2	1	1
CLOZSE	0	0	0	0	0	0	0	0	0
FIVFE	0	0	0	1	0	0	1	0	1
HARDT	6	4	6	5	3	5	7	5	3
HERDT	0	0	0	1	1	0	1	1	0
KILDT	0	0	0	0	0	1	0	0	1
LEAGK	0	0	0	0	0	0	0	0	0
RIDTER	1	0	1	1	2	1	2	1	1*
ROBPE	1	0	0	0	2	1	4	1	4
STEEN	1	1	2	2	2	2	2	2	1
WONDT	0	0	0	0	1	0	0	1	1

*had 2 errors on training data

Table 21: Test set errors, from 20 words, for the eleven minimally distinct word pairs using one hidden unit and trained until error<0.0005.

Word Pair	ϵ=0.1			ϵ=0.25			ϵ=0.4		
	α			α			α		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CSHIP	1	2	1	1	3	1	3	1	1
CLOZSE	0	0	0	0	0	0	0	0	0
FIVFE	0	1	1	0	0	0	0	1	1
HARDT	3	6	8	3	4	6	4	6	3
HERDT	0	1	1	2	2	0	1	1	1
KILDT	0	0	2	0	1	3	0	0	1
LEAGK	0	0	1	0	0	0	0	0	1
RIDTER	1	0	1	2	0	1	2	1	2
ROBPE	1	1	2	1	1	2	1	0	1
STEEN	1	1	1	1	1	2	1	1	2
WONDT	0	1	0	0	0	0	0	1	0

Table 22: Test set errors, from 20 words, for the eleven minimally distinct word pairs using one hidden unit and trained until error<0.00025.

Word Pair	ϵ=0.1			ϵ=0.25			ϵ=0.4		
	α			α			α		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CSHIP	0	1	1	2	1	3	1	2	3
CLOZSE	0	0	0	0	0	0	0	0	0
FIVFE	0	1	1	0	0	0	1	1	2
HARDT	1	0	1	3	2	2	4	2	2
HERDT	0	0	0	0	0	1	0	1	3
KILDT	0	0	0	0	0	0	0	0	0
LEAGK	0	0	0	0	0	0	0	0	0
RIDTER	0	0	0	0	0	0	0	1	0
ROBPE	0	0	0	0	0	0	1	0	0
STEEN	2	1	1	1	1	2	1	1	2
WONDT	1	0	1	1	0	1	0	0	1

Table 23: Test set errors, from 20 words, for the eleven minimally distinct word pairs using four hidden units and trained until error<0.0005.

Word Pair	ϵ=0.1			ϵ=0.25			ϵ=0.4		
	α			α			α		
	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CSHIP	1	1	1	1	1	2	1	1	2
CLOZSE	0	0	0	0	0	0	0	0	0
FIVFE	1	0	0	0	0	1	0	0	1
HARDT	0	1	1	1	1	1	2	0	2
HERDT	1	1	0	0	0	1	0	0	1
KILDT	1	1	0	1	0	0	1	0	1
LEAGK	0	0	0	0	0	0	0	0	0
RIDTER	1	0	0	1	0	0	0	0	0
ROBPE	0	0	0	0	0	0	0	0	0
STEEN	1	1	1	1	2	3	1	3	1
WONDT	0	0	0	1	1	0	1	0	1

Table 24: Test set errors, from 20 words, for the eleven minimally distinct word pairs using four hidden units and trained until error<0.00025.

Appendix D Graphs of Overall Energy Versus Length for all Eleven Word Pairs

These graphs show plots of overall energy versus length for all the minimal pairs considered in this memorandum. Note that the complete data set of forty words is shown in each graph.

The lines used for discrimination are shown.

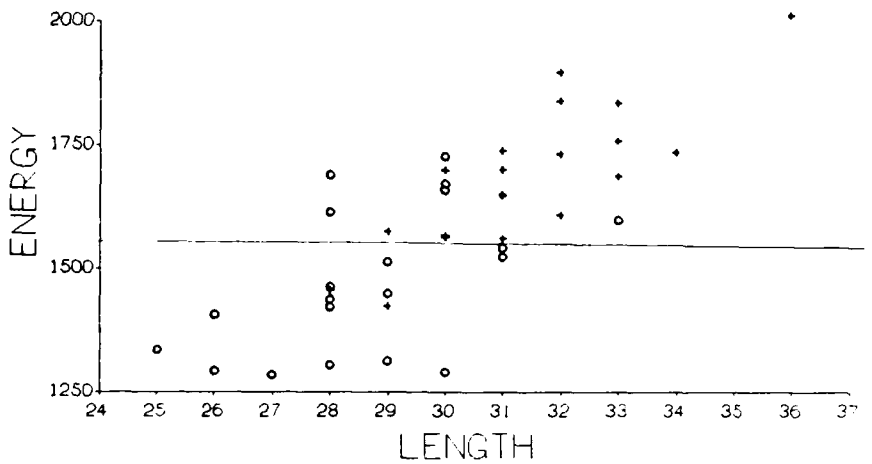


Figure 10: Overall energy versus length for the minimal pair CSHIP. Circles represent data from word class "chip", and crosses represent data from word class "ship".

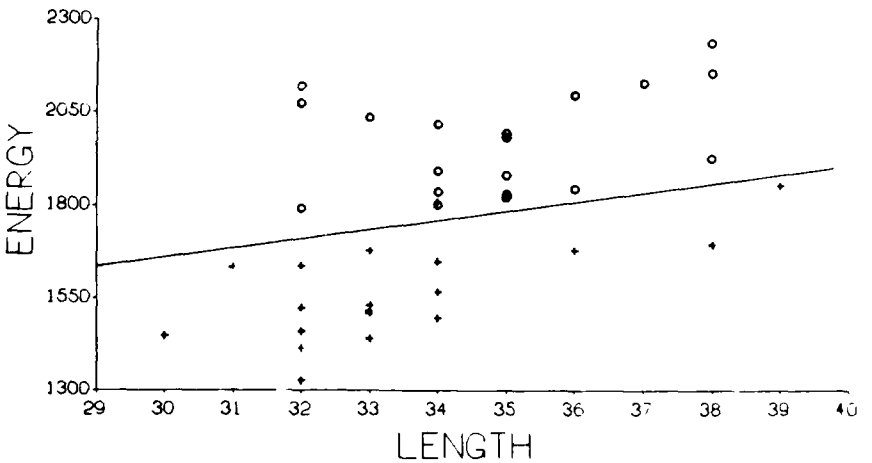


Figure 11: Overall energy versus length for the minimal pair CLOZSE. Circles represent data from word class "cloze", and crosses represent data from word class "close".

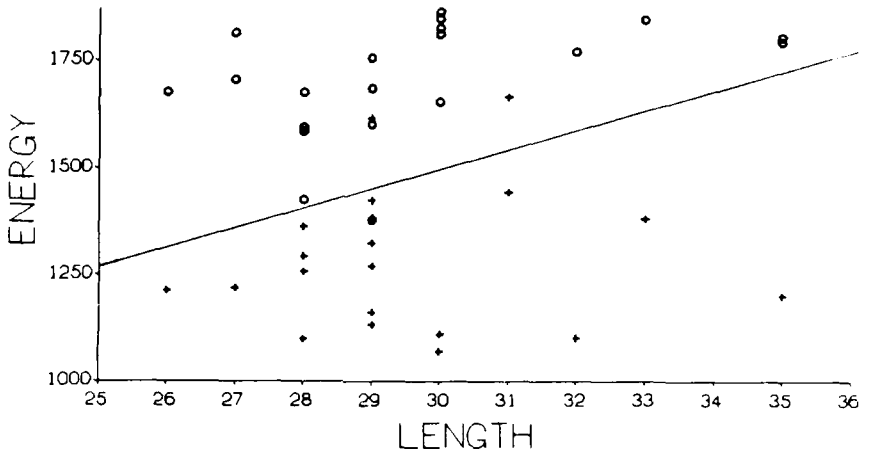


Figure 12: Overall energy versus length for the minimal pair FIVFE. Circles represent data from word class "five", and crosses represent data from word class "fife".

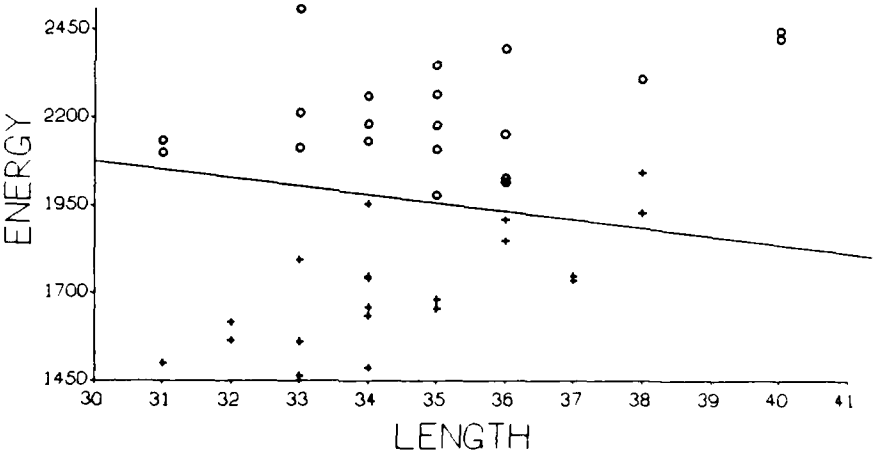


Figure 13: Overall energy versus length for the minimal pair HARDT. Circles represent data from word class "hard", and crosses represent data from word class "heart".

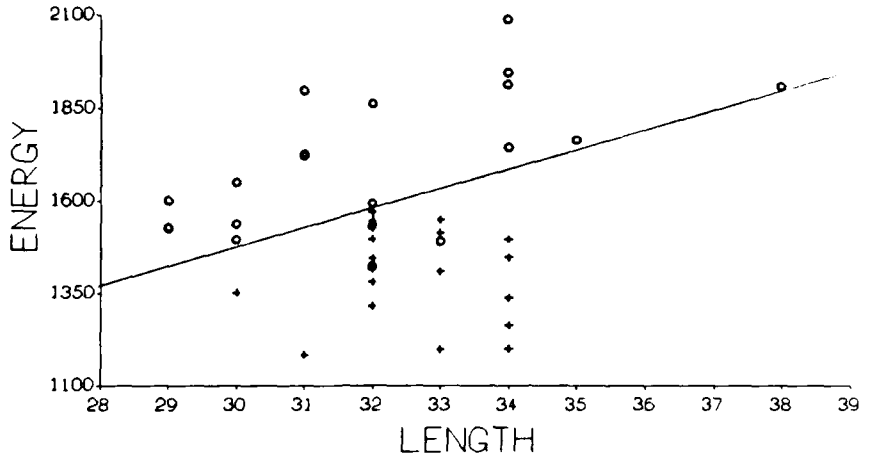


Figure 14: Overall energy versus length for the minimal pair HERDT. Circles represent data from word class "heard", and crosses represent data from word class "hurt".

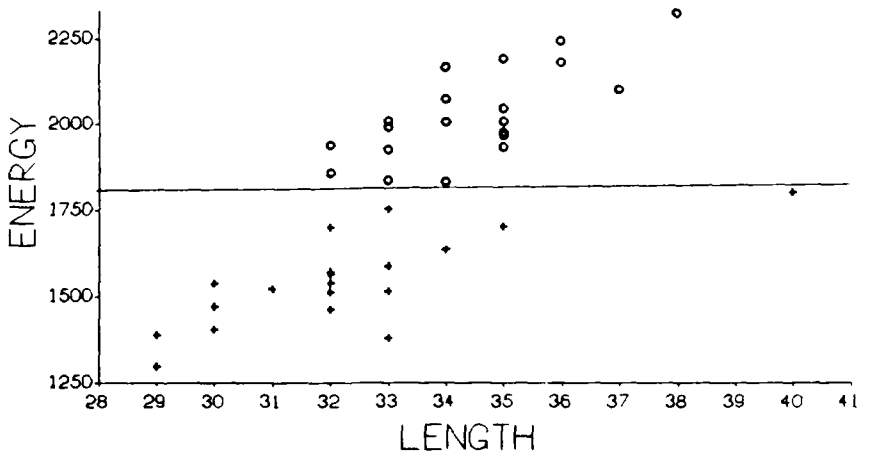


Figure 15: Overall energy versus length for the minimal pair KILDT. Circles represent data from word class "killed", and crosses represent data from word class "kilt".

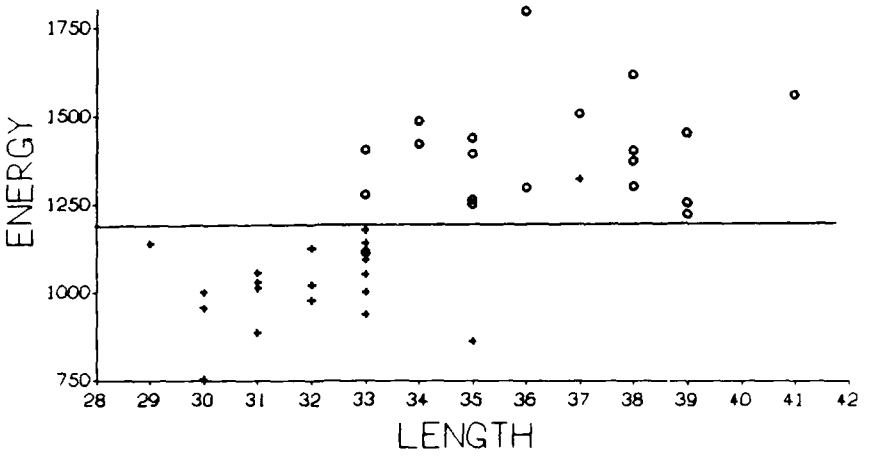


Figure 16: Overall energy versus length for the minimal pair LEEGK. Circles represent data from word class "league", and crosses represent data from word class "leak".

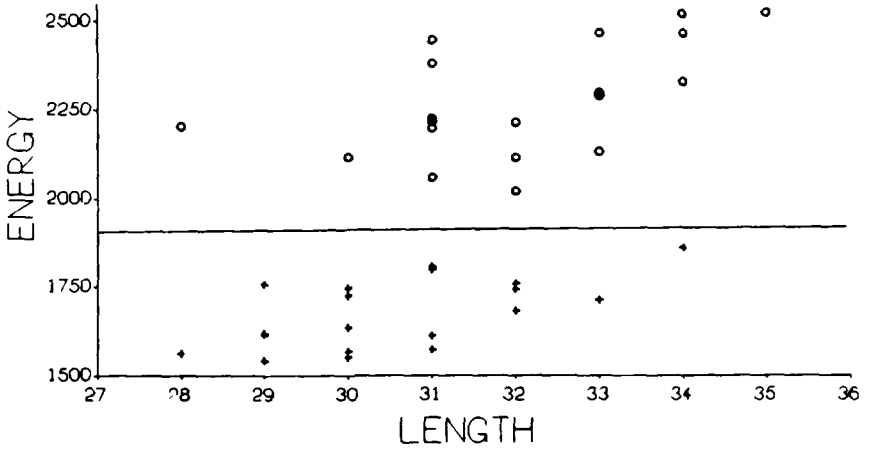


Figure 17: Overall energy versus length for the minimal pair RIDTER. Circles represent data from word class "rider", and crosses represent data from word class "writer".

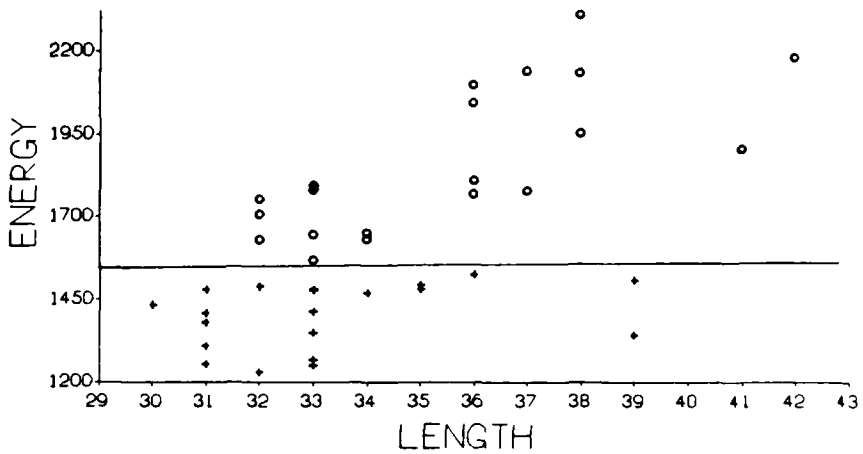


Figure 18: Overall energy versus length for the minimal pair ROBPE. Circles represent data from word class "robe", and crosses represent data from word class "rope".

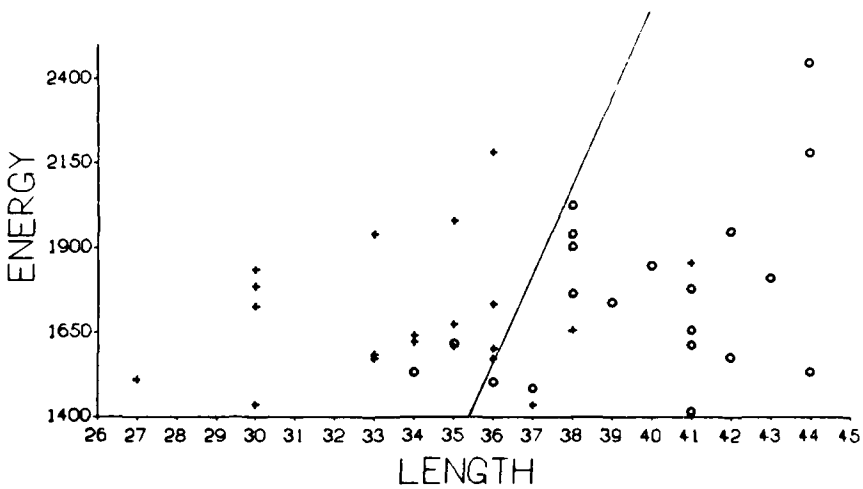


Figure 19: Overall energy versus length for the minimal pair STEEN. Circles represent data from word class "seen", and crosses represent data from word class "teen".

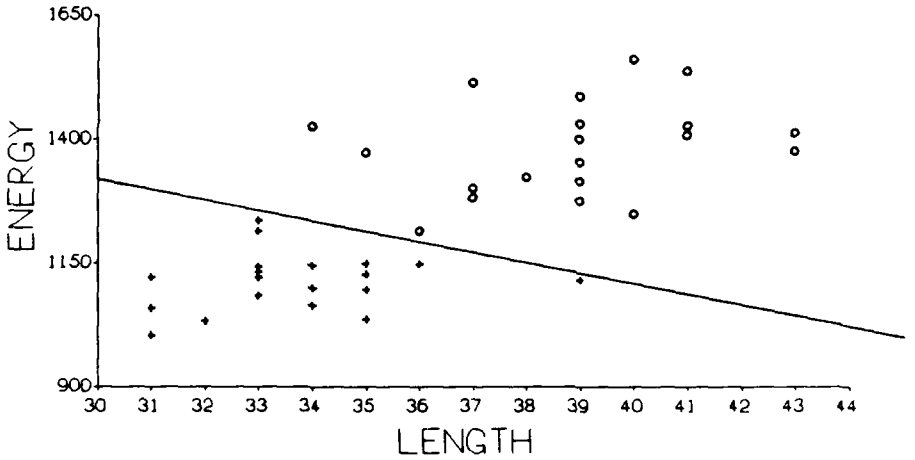


Figure 20: Overall energy versus length for the minimal pair WONDT. Circles represent data from word class "wand", and crosses represent data from word class "want".

DOCUMENT CONTROL SHEET

Overall security classification of sheet ..... **UNCLASSIFIED** .....

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the box concerned must be marked to indicate the classification eg (R) (C) or (S))

1. DRIC Reference (if known)	2. Originator's Reference	3. Agency Reference	4. Report Security Classification
	MEMO 4153		U/C

5. Originator's Code (if known)	6. Originator (Corporate Author) Name and Location
778400	Royal Signals and Radar Establishment St Andrews Road, Malvern, Worcs. WR14 3PS
5a. Sponsoring Agency's Code (if known)	6a. Sponsoring Agency (Contract Authority) Name and Location

7. Title

EXPERIMENTS IN MINIMALLY DISTINCT WORD-PAIR DISCRIMINATION USING THE MULTI-LAYER PERCEPTRON

7a. Title in Foreign Language (in the case of translations)

7b. Presented at (for conference papers) Title, place and date of conference

8. Author 1 Surname, initials	9(a) Author 2	9(b) Authors 3,4...	10. Date	pp. ref.
MOORE, R.K.	PEELING, S.M.		1988.4	3

11. Contract Number	12. Period	13. Project	14. Other Reference

15. Distribution statement

UNLIMITED

Descriptors (or keywords)

continue on separate piece of paper

Abstract

The *multi-layer perceptron* is investigated as a new approach to the automatic discrimination of spoken minimally distinct word-pairs. The choice of the parameters for the multi-layer perceptron is discussed and experimental results are reported. A comparison is made with hidden Markov modelling applied to the same data. The results, for this particular task, show that the discrimination accuracy obtained using the multi-layer perceptron is superior to that attained using hidden Markov modelling.